

New Perspectives on k -Support and Cluster Norms

Andrew M. McDonald, Massimiliano Pontil, Dimitris Stamos

Department of Computer Science
University College London
Gower Street, London WC1E
England, UK

E-mail: $\{a.mcdonald, m.pontil, d.stamos\}@cs.ucl.ac.uk$

Abstract

The k -support norm is a regularizer which has been successfully applied to sparse vector prediction problems. We show that it belongs to a general class of norms which can be formulated as a parameterized infimum over quadratics. We further extend the k -support norm to matrices, and we observe that it is a special case of the matrix cluster norm. Using this formulation we derive an efficient algorithm to compute the proximity operator of both norms. This improves upon the standard algorithm for the k -support norm and allows us to apply proximal gradient methods to the cluster norm. We also describe how to solve regularization problems which employ centered versions of these norms. Finally, we apply the matrix regularizers to different matrix completion and multitask learning datasets. Our results indicate that the spectral k -support norm and the cluster norm give state of the art performance on these problems, significantly outperforming trace norm and elastic net penalties.

Keywords: k -Support Norm, Proximal Methods, Regularization, Infimum Convolution, Matrix Completion.

1 Introduction

We study a family of norms that can be used as regularizers in vector and matrix learning problems. The norm is obtained by taking an infimum of certain quadratic functions, which are parameterized by a set Θ . By varying the set, the regularizer can be tailored to assumptions on the underlying model, which should lead to more accurate learning. The norm is defined for $w \in \mathbb{R}^d$, as

$$\|w\|_{\Theta} = \sqrt{\inf_{\theta \in \Theta} \sum_{i=1}^d \frac{w_i^2}{\theta_i}} \quad (1)$$

where Θ is a convex bounded subset of the positive orthant. This family is sufficiently rich to encompass standard regularizers such as the ℓ_p norms [Micchelli and Pontil, 2005] for $p \in [1, 2]$, the group Lasso Yuan and Lin [2006], Group Lasso with Overlap [Jacob et al., 2009b], the norm in [Jacob et al., 2009a], and the structured sparsity norms of [Micchelli et al., 2013]. Our work builds upon a recent line of papers which considered convex regularizers defined as an infimum problem over a parametric family of quadratics, as well as related infimal convolution problems [see Jacob et al., 2009b, Maurer and Pontil, 2012, Obozinski and Bach, 2012, and references therein].

In this paper, we focus on the specific case

$$\Theta = \left\{ \theta \in [a, b]^d : \sum_{i=1}^d \theta_i \leq c \right\}, \quad (2)$$

for parameters $0 < a < b$ and $c \in [da, db]$, and we refer to the corresponding norm as the box Θ -norm. As we show, this seemingly simple class nonetheless includes several nontrivial norms such as the recently proposed k -support norm [Argyriou et al., 2012]. This norm is interesting and important as it provides a tight convex relaxation of the cardinality operator on the ℓ_2 unit ball, and [Argyriou et al., 2012] demonstrated that it outperforms the Lasso and the elastic net [Zou and Hastie, 2005] on certain data sets. We derive an algorithm to compute the proximity operator of the squared box Θ norm which takes $\mathcal{O}(d \log d)$ time complexity, improving upon the $\mathcal{O}(d(k + \log d))$ complexity in the original work.

A second important aim of the paper is to extend our formulation to matrix learning regularization, in which the vector norm is applied to the spectrum of a matrix. Following a classical result by von Neumann [Von Neumann, 1937], we observe that if the set Θ is invariant under permutations then the formulation (1) can be extended to a orthogonally invariant matrix norm. In particular, this allows us to extend the k -support norm to a matrix setting. In this form we show that it is a special case of the *cluster norm* introduced by [Jacob et al., 2009a] in the context of clustered multitask learning. We also describe how to derive the proximity operator of the cluster norm, which enables efficient proximal gradient methods to be applied to this type of matrix learning problems.

In summary, the principal contributions of this paper are:

- We describe an $\mathcal{O}(d \log d)$ method to compute the box Θ -norm and its proximity operator, which allows us to employ optimal first order methods to solve the associated regularization problem;

- We show that the k -support norm is a box Θ -norm and observe that our algorithm to compute the proximity operator improves upon the $\mathcal{O}(d(k + \log d))$ algorithm by [Argyriou et al., 2012];
- We extend the norm (1) to the matrix setting, showing in particular that the spectral k -support spectral norm is a special case of the cluster norm.

On the experimental side, we apply these spectral Θ -norms to synthetic and real matrix learning datasets. Our findings indicate that the spectral k -support norm and the box Θ -norm produce state-of-the-art results on the matrix completion benchmarks. Furthermore we explain how to perform regularization with centered versions of these norms and we report state-of-the-art results on a multitask learning dataset in this setting. Finally we demonstrate numerically the efficiency of our algorithm to compute the proximity operator.

The paper is organized in the following manner. In Section 2, we summarize the properties of the Θ -norm and derive the k -support norm formulation. In Section 3.1, we compute the norm and in Section 4, we compute the proximity operator of the norm, including the k -support norm as a special case. In Section 5, we define the matrix norm and show the cluster norm coincides with the Θ -norm (with a particular choice of parameters a, b, c), and in Section 6, we report experiments on simulated and real data sets. Finally, in Section 7, we conclude. Derivations that are sketched or missing from the body of the paper are contained in the supplemental material.

2 Properties of the Norm

In this section, we highlight some important properties of the class of Θ -norms. We then specialize our observations to the box Θ -norms, showing in particular that it subsumes the k -support norm.

Note that the objective function of problem (1) is strictly convex, hence every minimizing sequence converges to the same point. The infimum is, however, not attained in general because a minimizing sequence may converge to a point on the boundary of Θ ; for instance, if $\Theta = \{\theta > 0 : \sum_{i=1}^d \theta_i \leq 1\}$, the minimizer is equal to the vector formed by the absolute values of the components of w . If Θ is closed, as in the case of (2), then the minimum is attained.

Our first result establishes that (1) is indeed a norm and derives the form of the dual norm.

Theorem 2.1. *If Θ is a convex bounded subset of the positive positive orthant in \mathbb{R}^d , then the infimum in equation (1) defines a norm and the dual norm is given by the formula*

$$\|u\|_{*,\Theta} = \sqrt{\sup_{\theta \in \Theta} \sum_{i=1}^d \theta_i u_i^2}. \quad (3)$$

Proof. (Sketch)¹ The objective in (3) is zero only if u is zero, absolutely homogeneous of order one and convex, so defines a norm. Using the fact that the Fenchel conjugate of $\frac{1}{2}\|\cdot\|^2$ is

¹Full derivations can be found in the appendix.

$\frac{1}{2}\|\cdot\|_*^2$ for any norm [see e.g. Boyd and Vandenberghe, 2004], a direct computation produces a separable optimization problem which can be solved explicitly to recover the form of the primal norm. ■

Note the symmetry between the expressions for the primal norm (1) and the dual norm (3). The latter is obtained by replacing the infimum with the supremum and the terms $1/\theta_i$ by θ_i . Note also that the computation of the dual norm involves searching for a supremum of a linear function of θ over the bounded convex set Θ . It follows that the supremum is achieved at an extreme point of the closure of Θ , that is, denoting by \mathcal{E} the set of such extreme points, the dual norm can be expressed as

$$\|u\|_{*,\Theta} = \sqrt{\max_{\theta \in \mathcal{E}} \sum_{i=1}^d \theta_i u_i^2}. \quad (4)$$

For the remainder of the paper, unless otherwise noted, we specialize our observations to the box Θ -norm, that is the case where Θ is defined as in (2). We assume throughout that $0 < a < b \leq 1$, which does not lose any generality since by rescaling and shifting a and b (with corresponding adjustment to c) we get an equivalent norm. We make the change of variable $\gamma_i = \frac{\theta_i - a}{b - a}$ and observe that the constraints on θ induce the constraint set $\{\gamma : 0 < \gamma_i \leq 1, \sum_{i=1}^d \gamma_i \leq \rho\}$, where $\rho = \frac{c - da}{b - a}$. Furthermore

$$\sum_{i=1}^d \theta_i u_i^2 = a\|u\|_2^2 + (b - a) \sum_{i=1}^d \gamma_i u_i^2.$$

We conclude that

$$\|u\|_{*,\Theta}^2 = a\|u\|_2^2 + (b - a) \sum_{j=1}^k (|u|_j^\downarrow)^2 + (\rho - k)|u_{k+1}|^2 \quad (5)$$

where $|u|^\downarrow$ is the vector obtained from u by reordering its components so that they are non-increasing in absolute value, and $k = \lfloor \rho \rfloor$. The expression simplifies if $\rho = k$, that is $c = (b - a)k + da$. This is exactly the setting considered by [Jacob et al., 2009a], where $k + 1$ is interpreted as the number of clusters and d as the number of tasks. We return to this interesting case in Section 5, where we explain in detail how to extend the Θ -norm to a matrix setting.

2.1 The k -Support Norm

We move on to discuss the connection between the box Θ -norms and the k -support norm. The latter was motivated by [Argyriou et al., 2012] as a tight convex relaxation of the cardinality operator on the ℓ_2 unit ball, and includes the ℓ_1 -norm and ℓ_2 -norm as special cases. Its dual is the ℓ_2 -norm of the largest k components, that is

$$\|u\|_{*,(k)} = \sqrt{\sum_{i=1}^k (|u_i|^\downarrow)^2} \quad (6)$$

where $k \leq d$. Comparing (6) with (5), we see that the k -support norm is obtained as a limiting case of the box Θ -norm for $a \rightarrow 0$, $b = 1$, and $c = k$.

Theorem 2.2. *If $\Theta = \{\theta \in \mathbb{R}^d : 0 < \theta_i \leq 1, \sum_{j=1}^d \theta_j \leq k\}$ then $\|w\|_{\Theta} = \|w\|_{(k)}$.*

[Argyriou et al., 2012] interpreted the k -support norm as a special case of the Group Lasso with Overlap introduced by [Jacob et al., 2009b], where the groups have support sets of cardinality at most k . The latter norm is defined as the infimal convolution

$$\|w\|_{\mathcal{G}} = \inf_{v \in \mathcal{V}(\mathcal{G})} \left\{ \sum_{g \in \mathcal{G}} \|v_g\|_2 : \sum_{g \in \mathcal{G}} v_g = w \right\} \quad (7)$$

where \mathcal{G} is a collection of subsets of $\{1, \dots, d\}$ and

$$\mathcal{V}(\mathcal{G}) = \{(v_g)_{g \in \mathcal{G}} : v_g \in \mathbb{R}^d, \text{supp}(v_g) \subseteq g, g \in \mathcal{G}\}.$$

The k -support norm corresponds to the case that \mathcal{G} is formed by all subsets containing at most k elements, that is $\mathcal{G} = \mathcal{G}_k := \{g : g \subseteq \{1, \dots, d\}, |g| \leq k\}$. Furthermore, as noted by [Jacob et al., 2009b], the dual norm is given by

$$\sqrt{\max_{g \in \mathcal{G}} \sum_{i \in g} u_i^2}.$$

Writing $\sum_{g \in \mathcal{G}} u_i^2 = \sum_{i=1}^d (1_g)_i u_i^2$, where $(1_g)_i = 1$ if $i \in g$ and zero otherwise, we see that the norm (7) is a (general) Θ -norm, where Θ is the convex hull of $\{1_g : g \in \mathcal{G}\}$.

We end this section with an infimal convolution interpretation of the box Θ -norm.

Proposition 2.1. *If Θ is of the form in equation (2) and $c = (b - a)k + da$, for $k \in \{1, \dots, d\}$, then*

$$\|w\|_{\Theta} = \inf \left\{ \sum_{g \in \mathcal{G}_k} \sqrt{\sum_{i \in g} \frac{v_{g,i}^2}{b} + \sum_{i \notin g} \frac{v_{g,i}^2}{a}} : \sum_{g \in \mathcal{G}_k} v_g = w \right\}.$$

Note that if $b = 1$, then as a tends to zero, we obtain the expression of the k -support norm (7) with $\mathcal{G} = \mathcal{G}_k$, recovering in particular the support constraints. If a is small and positive, the support constraints are not imposed, however, most of the weight for each v_g tends to be concentrated on $\text{supp}(g)$. Hence, Proposition 2.1 suggests that the box Θ -norm regularizer will encourage vectors w whose dominant components are a subset of a union of a small number of groups $g \in \mathcal{G}_k$.

3 Computation of the Norm

In this section, we compute the box Θ -norm by explicitly solving the optimization problem (1).

Theorem 3.1. For $w \in \mathbb{R}^d$,

$$\|w\|_{\Theta}^2 = \frac{1}{b^2} \|w_Q\|_2^2 + \frac{1}{p} \|w_I\|_1^2 + \frac{1}{a^2} \|w_L\|_2^2, \quad (8)$$

where $w_Q = (|w|_1^\downarrow, \dots, |w|_q^\downarrow)$, $w_I = (|w|_{q+1}^\downarrow, \dots, |w|_{d-\ell}^\downarrow)$, $w_L = (|w|_{d-\ell+1}^\downarrow, \dots, |w|_d^\downarrow)$, q and ℓ are the unique integers in $\{0, \dots, d\}$, satisfying $q + \ell \leq d$ and

$$\frac{|w_q|}{b} \geq \frac{1}{p} \sum_{i=q+1}^{d-\ell} |w_i| > \frac{|w_{q+1}|}{b}, \quad (9)$$

$$\frac{|w_{d-\ell}|}{a} \geq \frac{1}{p} \sum_{i=q+1}^{d-\ell} |w_i| > \frac{|w_{d-\ell+1}|}{a}, \quad (10)$$

where $p = c - qb - \ell a$ and we have defined $|w_0| = \infty$ and $|w_{d+1}| = 0$.

Proof. (Sketch) We need to solve the optimization problem

$$\inf_{\theta} \left\{ \sum_{i=1}^d \frac{w_i^2}{\theta_i} : a \leq \theta_i \leq b, \sum_{i=1}^d \theta_i \leq c \right\}. \quad (11)$$

Without loss of generality we assume that the w_i are ordered non increasing in absolute values, and it follows that at the optimum the θ_i are also ordered non increasing (see Lemma A.2). We further assume without loss of generality that $w_i \neq 0$ for all i and $c \leq db$, so the sum constraint will be tight at the optimum (see Remark A.1).

The Lagrangian is given by

$$L(\theta, \alpha) = \sum_{i=1}^d \frac{w_i^2}{\theta_i} + \frac{1}{\alpha^2} \left(\sum_{i=1}^d \theta_i - c \right)$$

where $1/\alpha^2$ is a strictly positive multiplier to be chosen such that $S(\alpha) := \sum_{i=1}^d \theta_i = c$. We can then solve the original problem by minimizing the Lagrangian over the constraint $\theta \in [a, b]^d$. Due to the coupling effect of the multiplier, we can solve the simplified problem componentwise using Lemma A.3, obtaining the solution

$$\theta_i = \begin{cases} a, & \text{if } \alpha < \frac{a}{|w_i|}, \\ \alpha |w_i|, & \text{if } \frac{a}{|w_i|} \leq \alpha \leq \frac{b}{|w_i|}, \\ b, & \text{if } \alpha > \frac{b}{|w_i|}, \end{cases} \quad (12)$$

where $S(\alpha) = c$. The minimizer has the form $\theta = (b, \dots, b, \theta_{q+1}, \dots, \theta_{d-\ell}, a, \dots, a)$, where q, ℓ are determined by the value of α . From $S(\alpha) = c$ we get $\alpha = p / (\sum_{i=q+1}^{d-\ell} |w_i|)$. The value of the norm in (8) follows by substituting θ into the objective. Finally, by construction we have $\theta_q \geq b > \theta_{q+1}$ and $\theta_{d-\ell} > a \geq \theta_{d-\ell+1}$, implying (9) and (10). ■

Theorem 3.1 suggests two methods for computing the box Θ -norm. First we find α such that $S(\alpha) = c$; this value uniquely determines θ in (12), and the norm follows by substitution into (11). Alternatively we identify q and ℓ that jointly satisfy (9) and (10) and we compute the norm using (8). Taking advantage of the structure of θ we now show that the first method leads to a computation time that is $\mathcal{O}(d \log d)$.

Theorem 3.2. *The computation of the Θ -norm can be completed in $\mathcal{O}(d \log d)$ time.*

Proof. Following Theorem 3.1, we need to choose α^* to satisfy the coupling constraint $S(\alpha^*) = c$. Each component of θ is a piecewise linear function in the form of a step function with a constant slope between the values a and b . Let the set $\{\alpha^i\}_{i=1}^{2d}$ be the set of the $2d$ critical points, where the α^i are ordered non decreasing. The function $S(\alpha)$ is a non decreasing piecewise linear function with at most $2d$ critical points.

We can find α^* by first sorting the points $\{\alpha^i\}$, finding α^i and α^{i+1} such that $S(\alpha^i) \leq c \leq S(\alpha^{i+1})$ by binary search, and then interpolating α^* between the two points.

Sorting takes $\mathcal{O}(d \log d)$. Computing $S(\alpha^i)$ at each step of the binary search is $\mathcal{O}(d)$, so $\mathcal{O}(d \log d)$ overall. Given α^i and α^{i+1} , interpolating α^* is $\mathcal{O}(d)$, so the algorithm overall is $\mathcal{O}(d \log d)$ as claimed. ■

As outlined earlier, the k -support norm is a special case of the box Θ -norm. We have the following corollary of Theorem 3.1 and Theorem 3.2.

Corollary 3.1. *For $w \in \mathbb{R}^d$, and $k \leq d$,*

$$\|w\|_{(k)} = \left(\sum_{j=1}^q |w_j^\downarrow|^2 + \frac{1}{k-q} \left(\sum_{j=q+1}^d |w_j^\downarrow| \right)^2 \right)^{\frac{1}{2}},$$

where q is the unique integer in $\{0, k-1\}$ satisfying

$$|w_q| \geq \frac{1}{k-q} \sum_{j=q+1}^d |w_j| > |w_{q+1}|, \quad (13)$$

and we have defined $w_0 = \infty$. Furthermore, the norm can be computed in $\mathcal{O}(d \log d)$ time.

Proof. The first claim follows from Theorem 3.1 by letting $a \rightarrow 0$, $b = 1$, $c = k$ and noting that equation (10) is no longer required. The complexity follows from Theorem 3.2. ■

4 Proximity Operator

Proximal gradient methods can be used to efficiently solve optimization problems of the form $\min_{w \in \mathbb{R}^d} f(w) + \lambda g(w)$, where f is a convex loss function with Lipschitz continuous gradient, $\lambda > 0$ is a regularization parameter, and g is a convex function for which the proximity operator can be computed efficiently [see Combettes and Pesquet, 2011, Beck and Teboulle, 2009,

Nesterov, 2007, and references therein]. The proximity operator of g with parameter $\rho > 0$ is defined as

$$\text{prox}_{\rho g}(w) = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2} \|x - w\|^2 + \rho g(x) \right\}.$$

We now use the infimum formulation of the box Θ -norm to derive the proximity operator of the squared norm.

Theorem 4.1. *The proximity operator of the square of the box Θ -norm at point $w \in \mathbb{R}^d$ with parameter $\frac{\lambda}{2}$ is given by $\text{prox}_{\frac{\lambda}{2} \|\cdot\|_{\Theta}^2}(w) = (\frac{\theta_1 w_1}{\theta_1 + \lambda}, \dots, \frac{\theta_d w_d}{\theta_d + \lambda})$, where*

$$\theta_i = \begin{cases} a, & \text{if } \alpha < \frac{a+\lambda}{|w_i|}, \\ \alpha|w_i| - \lambda, & \text{if } \frac{a+\lambda}{|w_i|} \leq \alpha \leq \frac{b+\lambda}{|w_i|} \\ b, & \text{if } \alpha > \frac{b+\lambda}{|w_i|}, \end{cases} \quad (14)$$

and α is chosen such that $S(\alpha) = \sum_{i=1}^d \theta_i = c$.

Proof. Similarly to Remark A.1 for Theorem 3.1, we can assume without loss of generality that $w_i \neq 0$. Using the infimum formulation of the norm, and noting that the infimum is attained, we solve

$$\min_{x \in \mathbb{R}^d} \min_{\theta \in \Theta} \left\{ \frac{\lambda}{2} \sum_{i=1}^d \frac{x_i^2}{\theta_i} + \frac{1}{2} \sum_{i=1}^d (x_i - w_i)^2 \right\}.$$

We can exchange the order of the optimization and solve for x first. The problem is separable we find $x_i = \frac{\theta_i w_i}{\theta_i + \lambda}$. Discarding a multiplicative factor of $\lambda/2$, the problem in θ becomes

$$\min_{\theta} \left\{ \sum_{i=1}^d \frac{w_i^2}{\theta_i + \lambda} : a \leq \theta_i \leq b, \sum_{i=1}^d \theta_i \leq c \right\}$$

and we can equivalently solve

$$\min_{a \leq \theta_i \leq b} \left\{ \sum_{i=1}^d \frac{w_i^2}{\theta_i + \lambda} + \frac{1}{\alpha^2} \left(\sum_{i=1}^d \theta_i - c \right)^2 \right\},$$

where $\frac{1}{\alpha^2}$ is a strictly positive Lagrange multiplier and α is chosen to ensure that $S(\alpha) = c$. The problem is separable and decouples, and by Lemma A.3 the solution is given by

$$\theta_i = \begin{cases} a, & \text{if } \alpha < \frac{a+\lambda}{|w_i|}, \\ \alpha|w_i| - \lambda, & \text{if } \frac{a+\lambda}{|w_i|} \leq \alpha \leq \frac{b+\lambda}{|w_i|} \\ b, & \text{if } \alpha > \frac{b+\lambda}{|w_i|}, \end{cases}$$

where α is chosen such that $S(\alpha) = c$. ■

By the same reasoning as for Theorem 3.2 we have the following.

Theorem 4.2. *The computation of the proximity operator can be completed in $\mathcal{O}(d \log d)$ time.*

Algorithm 1 illustrates the computation of the proximity operator for the squared box Θ -norm in $\mathcal{O}(d \log d)$ time. This includes the k -support as a special case, which improves upon the complexity of the $\mathcal{O}(d(k + \log d))$ computation provided in [Argyriou et al., 2012]. We summarize this in the following corollary.

Corollary 4.1. *The proximity operator of the square of the k -support norm at point w with parameter $\frac{\lambda}{2}$ is given by $\text{prox}_{\frac{\lambda}{2} \|\cdot\|_{\Theta}^2}(w) = x$, where $x_i = \frac{\theta_i w_i}{\theta_i + \lambda}$, and*

$$\theta_i = \begin{cases} 0, & \text{if } \alpha < \frac{\lambda}{|w_i|}, \\ \alpha |w_i| - \lambda, & \text{if } \frac{\lambda}{|w_i|} \leq \alpha \leq \frac{1+\lambda}{|w_i|} \\ 1, & \text{if } \alpha > \frac{1+\lambda}{|w_i|}, \end{cases}$$

where α is chosen such that $S(\alpha) = k$. Furthermore, the proximity operator can be computed in $\mathcal{O}(d \log d)$ time.

Proof. The expression follows directly from Theorem 4.1 by letting $a \rightarrow 0$, and setting $b = 1$ and $c = k$ in equation (14) and the complexity claim follows directly from Theorem 4.2. ■

Algorithm 1 Computation of $x = \text{prox}_{\frac{\lambda}{2} \|\cdot\|_{\Theta}^2}(w)$.

Require: parameters a, b, c, λ .

1. Sort points $\{\alpha^i\}_{i=1}^d = \left\{ \frac{a+\lambda}{|w_j|}, \frac{b+\lambda}{|w_j|} \right\}_{j=1}^d$ such that $\alpha^i \leq \alpha^{i+1}$.
 2. Identify points α^i and α^{i+1} such that $S(\alpha^i) \leq c$ and $S(\alpha^{i+1}) \geq c$ by binary search.
 3. Find α^* between α^i and α^{i+1} such that $S(\alpha^*) = c$ by linear interpolation.
 4. Compute $\theta_i(\alpha^*)$ for $i = 1 \dots, d$.
 5. Return $x_i = \frac{\theta_i w_i}{\theta_i + \lambda}$ for $i = 1 \dots, d$.
-

5 Spectral Regularization

In this section, we use the box Θ -norm to define a class of orthogonally invariant matrix norms. In particular, we extend the k -support norm to this setting and note that it is closely related to the cluster norm of [Jacob et al., 2009a].

A norm $\|\cdot\|$ on $\mathbb{R}^{d \times m}$ is called orthogonally invariant if

$$\|W\| = \|UWV^\top\|, \quad (15)$$

for any orthogonal matrices $U \in \mathbb{R}^{d \times d}$ and $V \in \mathbb{R}^{m \times m}$. A classical result of von Neumann [Von Neumann, 1937] establishes that a norm is orthogonal invariant if and only if it is of the

form $\|W\| = g(\sigma(W))$, where $\sigma(W)$ is the vector formed by the singular values of W , and g is a *symmetric gauge function*. This means that g is a norm which is invariant under permutations and sign changes. That is, $g(w_1, \dots, w_d) = g(w_{\pi(1)}, \dots, w_{\pi(d)})$ for every permutation π , and $g(Jw) = g(w)$ for every diagonal matrix J with entries ± 1 .

Considering the box Θ -norm objective in equation (1), we note that it is sign invariant. Moreover, if the set Θ is permutation invariant (meaning that if $\theta \in \Theta$ then $\theta_\pi \in \Theta$ for every permutation π), then the Θ -norm is permutation invariant. This holds for the set (2) as the components θ_i each lie in the same interval $[a, b]$, and the sum is invariant to the order of the terms by Lemma A.2, we conclude that the corresponding box Θ -norm is a symmetric gauge function. By applying this norm to the spectrum of a matrix we obtain an orthogonally invariant norm which we term a *spectral Θ -norm*. In particular choosing $a = 0, b = 1, c = k$, we obtain the spectral k -support norm, which we define (with some abuse of notation) as $\|W\|_{(k)} = \|\sigma(W)\|_{(k)}$.

5.1 Cluster Norm for Multi Task Learning

We now briefly discuss multitask learning. The problem is given by

$$\min_{W \in \mathbb{R}^{d \times m}} \ell(W) + \lambda \Omega(W)$$

for some loss function ℓ and regularizer Ω . The columns of W represent different regression vectors (tasks). A natural assumption that arises in applications is that the tasks are clustered and a matrix regularizer can be chosen to favour such structure. We refer to [Evgeniou et al., 2005, Argyriou et al., 2008, Jacob et al., 2009a] for a discussion and motivating examples.

In particular, [Jacob et al., 2009a] propose a composite regularizer which balances a tradeoff between the norm of the mean of the tasks, the variance between the clusters, and the variance within the clusters and define the cluster norm as

$$\|W\|_{\text{cl}} = \sqrt{\inf_{\Sigma \in \mathcal{S}} \text{tr}(W \Sigma^{-1} W^\top)} \quad (16)$$

where \mathcal{S} is a set of symmetric positive definite matrices subject to a spectral constraint,

$$\mathcal{S} = \{\Sigma \succeq 0 : aI \preceq \Sigma \preceq bI, \text{tr} \Sigma = c\} \quad (17)$$

and $c = (b - a)k + ma$. Here $k + 1$ plays the role of the number of clusters and the parameters a and b control the variances within and between the clusters respectively [see Jacob et al., 2009a, for a discussion].

The cluster norm corresponds to the matrix norm where we apply the box Θ -norm to the singular values of the matrix.

Proposition 5.1. *The cluster norm (16) is an orthogonally invariant norm and can be written as*

$$\|W\|_{\text{cl}} = \|\sigma(W)\|_{\Theta} = \sqrt{\inf_{\theta \in \Theta} \sum_{i=1}^d \frac{\sigma_i(W)^2}{\theta_i}}, \quad (18)$$

where $\Theta = \left\{ \theta \in \mathbb{R}^d : a \leq \theta \leq b, \sum_{i=1}^d \theta_i \leq c \right\}$.

Proof. Let (U, Σ, V) be the singular value decomposition of W and note that

$$\begin{aligned}\text{tr}(WE^{-1}W^\top) &= \text{tr}(U\Sigma V^\top E^{-1}V\Sigma^\top U^\top) \\ &= \text{tr}(\Sigma V^\top E^{-1}V\Sigma^\top),\end{aligned}$$

where the second equality follows by cyclic property of the trace operator. Now let $E' = V^\top E^{-1}V$ and observe that $\sigma(E) = \sigma(E')$. We conclude that the cluster norm is orthogonally invariant and can be expressed in the form (18). ■

The computational considerations in Sections 3 and 4 can be naturally extended to the matrix setting by using von Neumann's trace inequality [see e.g. Horn and Johnson, 1991, ex. 3.3.10]. Here we only comment on the computation of the proximity operator which is important for our numerical experiments below. The proximity operator of an orthogonally invariant norm $\|\cdot\| = g(\sigma(\cdot))$ is given by

$$\text{prox}_{\|\cdot\|}(W) = U \text{diag}(\text{prox}_{\|\cdot\|}(\sigma(W))) V^\top,$$

where U and V are the matrices formed by the left and right singular vectors of W [see, e.g. Argyriou et al., 2011, Prop 3.1]. Using this result we can apply any spectral Θ -norm, such as the spectral k -support norm, to matrix learning problems as we describe in Section 6 below.

5.2 Centered Cluster Norm

In their work, [Jacob et al., 2009a] consider the following regularization problem:

$$\min_{W \in \mathbb{R}^{d \times m}} \ell_c(W) + \lambda \|W\Pi\|_{\text{cl}}^2. \quad (19)$$

where $\Pi = I - \frac{11^\top}{m}$, is the centering operator, so that $\Pi W = [w_1 - \bar{w}, \dots, w_m - \bar{w}]$ with $\bar{w} = \frac{1}{m} \sum_{i=1}^m w_i$ and ℓ_c is a modified loss term which incorporates the square norm regularization of the mean of the tasks. We end this section by discussing how to solve the optimization problem (19). The following lemma is key.

Lemma 5.1. *For all $W = [w_1, \dots, w_m] \in \mathbb{R}^{d \times m}$, it holds*

$$\|W\Pi\|_{\text{cl}} = \min_{z \in \mathbb{R}^d} \|[w_1 - z, \dots, w_m - z]\|_{\text{cl}}.$$

Proof. Since $\|\cdot\|_{\text{cl}}$ is orthogonally invariant

$$\begin{aligned}\|W\Pi\|_{\text{cl}}^2 &= \|(W\Pi)^\top\|_{\text{cl}}^2 \\ &= \inf_{\Sigma \in \mathcal{S}_c} \text{tr}((W\Pi)^\top \Sigma^{-1} (W\Pi)) \\ &= \inf_{\Sigma \in \mathcal{S}_c} \sum_{i=1}^m (w_i - \bar{w})^\top \Sigma^{-1} (w_i - \bar{w}) \\ &= \inf_{\Sigma \in \mathcal{S}_c} \min_{z \in \mathbb{R}^d} \sum_{i=1}^m (w_i - z)^\top \Sigma^{-1} (w_i - z) \\ &= \min_{z \in \mathbb{R}^d} \|[w_1 - z, \dots, w_m - z]\|_{\text{cl}},\end{aligned}$$

where we have used the fact that $W\Pi$ is the centred weight matrix $[w_1 - \bar{w}, \dots, w_m - \bar{w}]$, with $\bar{w} = \frac{1}{m} \sum_{i=1}^m w_i$, and the fact that for $U \in \mathbb{R}^{p \times m}$, $S \in \mathbb{R}^{p \times q}$, $V \in \mathbb{R}^{q \times m}$, where U, V have columns u_i, v_i respectively, we have

$$\text{tr}(U^\top S V) = \sum_{i=1}^m u_i^\top S v_i.$$

Finally, note that the quadratic $\sum_{i=1}^m (w_i - z)^\top \Sigma^{-1} (w_i - z)$ is minimized at $z = \bar{w}$. \blacksquare

Using this lemma, and letting $v_i = w_i - z$, and $V = [w_1 - z, \dots, w_m - z]$, we rewrite problem (19) as the equivalent problem in $m(d+1)$ variables

$$\min_{(V, z) \in \mathbb{R}^{d \times m} \times \mathbb{R}^d} \ell_c([v_1 + z, \dots, v_m + z]) + \lambda \|V\|_{\text{cl}}^2. \quad (20)$$

This problem is of the form $\ell(V, z) + f(V, z)$, where $f(V, z) = \lambda \|V\|_{\text{cl}}^2$. Using this formulation, we can directly apply the proximal gradient method using the proximity operator computation for the cluster norm, since $\text{prox}_f(V_0, z_0) = (\text{prox}_{\lambda \|\cdot\|_{\text{cl}}}(V_0), z_0)$. Finally we point out that this method can be used to perform optimization with the centered trace norm and the centered spectral k -support norm since both can be written as a cluster norm.

6 Numerical Experiments

In this section, we report on the statistical performance of the spectral Θ -norms in regularization problems, and we briefly comment on the numerical performance of Algorithm 1 for the k -support norm. In [Argyriou et al., 2012], the authors demonstrated the good estimation properties of the vector k -support norm compared to the Lasso and the elastic net. Here, we investigate the spectral Θ -norms including the spectral k -support norm. For both matrix completion and multitask learning we consider the spectral k -support (ks) and the spectral Θ -norm (box). In the latter case we also consider the centered cluster norm ($c-cn$) and the centered spectral k -support norm ($c-ks$). For both set of experiments we included the trace norm [see, e.g. Cai et al., 2008] and the spectral elastic net.

All optimizations used an accelerated proximal gradient method (FISTA), [see e.g. Beck and Teboulle, 2009, Combettes and Pesquet, 2011, Nesterov, 2007].² We report on test errors, number of iterations to convergence (N), final matrix rank (r) and parameter values for k and a , all of which are chosen by validation along with the regularization parameter.

6.1 Proximity Operator

We first verified the improved performance of the proximity operator computation. Table 1 illustrates the time taken by the $\mathcal{O}(d(k + \log d))$ method proposed in [Argyriou et al., 2012] and by our $\mathcal{O}(d \log d)$ method on a synthetic noisy dataset, fixing $k = d/100$, and confirms the theoretical improvement.

²Code used in the experiments is available at <http://www0.cs.ucl.ac.uk/staff/M.Pontil/software.html>.

Table 1: Comparison of proximity operator algorithms for k -support norm (in s), $k = d/100$. Algorithm 1 is the method in [Argyriou et al., 2012], Algorithm 2 is our method.

k	10	20	40	80	160
Alg. 1	0.0101	0.0361	0.1339	0.5176	2.0298
Alg. 2	0.0012	0.0017	0.0028	0.0044	0.0096

6.2 Simulated Data

Matrix completion. We applied the regularizer to matrix completion on noisy observations of low rank matrices. Each $m \times m$ matrix is generated as $W = UV^T + \mathcal{E}$, where $U, V \in \mathbb{R}^{m \times r}$, $r \ll m$, and the entries of U, V and \mathcal{E} are i.i.d. standard Gaussian.

We sampled uniformly a percentage ρ of the entries for training, of which 10% were used in validation. The error was measured as $\frac{\|\text{true} - \text{predicted}\|^2}{\|\text{true}\|^2}$ [Mazumder et al., 2010] and averaged over 100 trials. The convergence criterion was the percentage change in objective, with a tolerance of 10^{-5} . The results are summarized in Table 2.

In all training regimes, the spectral Θ -norm generated the lowest test errors, and converged within the fewest iterations. While it generated a matrix with the highest rank, the spectrum shows a rapid decrease after the first few singular values, which suggests a thresholding approach might be beneficial. In the absence of noise, the spectral Θ -norm performed worse than the trace and spectral k -support norms, which suggests that the lower bound on the non zero singular values due to a counteracted the noise in this instance.

Clustered Learning. As a proof of concept we tested the centered cluster norm and the centered k -support norm on an synthetic dataset. We generated a 100×100 , rank 5, block diagonal matrix, where the entries of each 20×20 block were equal to a given constant (between 1 and 10) plus noise. Table 3 illustrates the results for different training set sizes averaged over 100 runs. The cluster norm performed the best, however it generated full rank matrices (with the characteristic rapid spectral decay). Furthermore, the centered spectral k -support norm was a close second, with test errors converging towards those of the cluster norm as the training set increased in size, while maintaining a low rank due to the sparsity-inducing properties of the k -support norm. Figure 1 illustrates a sample matrix along with the final matrix determined by the cluster and trace norms.

6.3 Matrix Completion

Next we considered different matrix completion data sets, where a percentage of the (user,rating) entries of a matrix are observed, and the task is to predict the unobserved ratings, with the assumption that the true underlying matrix is low rank. We considered the following datasets.

MovieLens 100k.³ This dataset consists of 943 users and 1,682 movies, with a total of 100,000 ratings from 1 to 5.

³<http://grouplens.org/datasets/movielens/>

Table 2: Matrix completion on simulated data sets.

dataset	norm	test error	N	r	k	a
rank 5 $\rho=10\%$	tr	0.8184 (0.03)	77	20	-	-
	en	0.8164 (0.03)	65	20	-	-
	ks	0.8036 (0.03)	39	16	3.6	-
	box	0.7805 (0.03)	33	87	2.9	0.017
rank 5 $\rho=15\%$	tr	0.5764 (0.04)	60	22	-	-
	en	0.5744 (0.04)	56	21	-	-
	ks	0.5659 (0.03)	42	18	3.3	-
	box	0.5525 (0.04)	30	100	1.3	0.009
rank 5 $\rho=20\%$	tr	0.4085 (0.03)	47	23	-	-
	en	0.4081 (0.03)	45	23	-	-
	ks	0.4031 (0.03)	33	21	3.1	-
	box	0.3898 (0.03)	25	100	1.3	0.009
rank 10 $\rho=20\%$	tr	0.6356 (0.03)	49	27	-	-
	en	0.6359 (0.03)	48	27	-	-
	ks	0.6284 (0.03)	34	24	4.4	-
	box	0.6243 (0.03)	25	89	1.8	0.009
rank 10 $\rho=30\%$	tr	0.3642 (0.02)	68	36	-	-
	en	0.3638 (0.02)	61	36	-	-
	ks	0.3579 (0.02)	42	33	5.0	-
	box	0.3486 (0.02)	30	100	2.5	0.009

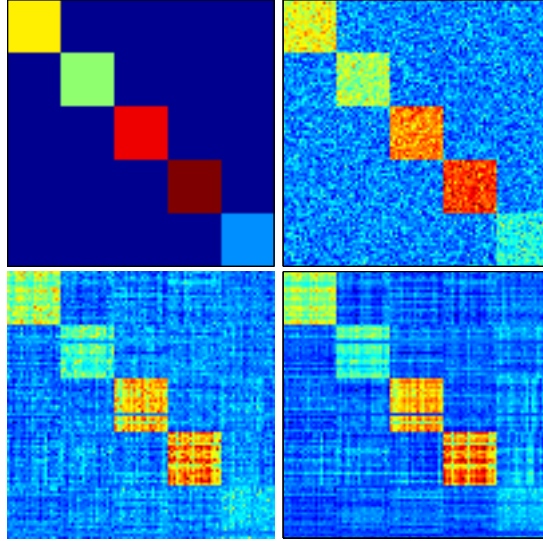


Figure 1: Block clustered matrix and recovered solution. Clockwise from top left: true, noisy, centered cluster norm, trace norm.

Table 3: Clustered block diagonal matrix.

dataset	norm	test error	r	k	a
$\rho=10\%$	tr	0.6909 (0.07)	23	-	-
	en	0.6876 (0.08)	22	-	-
	ks	0.6740 (0.08)	21	3.10	-
	box	0.6502 (0.07)	100	2.86	0.0052
	c-ks	0.6211 (0.05)	16	2.21	-
	c-cn	0.6085 (0.05)	100	2.28	0.002
$\rho=15\%$	tr	0.3845 (0.05)	22	-	-
	en	0.3831 (0.05)	22	-	-
	ks	0.3820 (0.05)	22	2.70	-
	abc	0.3798 (0.05)	100	2.48	0.004
	c-ks	0.3495 (0.04)	19	2.01	-
	c-cn	0.3446 (0.03)	100	2.10	0.001
$\rho=20\%$	tr	0.2596 (0.03)	25	-	-
	en	0.2592 (0.03)	25	-	-
	ks	0.2576 (0.03)	23	2.80	-
	box	0.2516 (0.03)	100	2.53	0.006
	c-ks	0.2228 (0.02)	22	2.26	-
	c-cn	0.2221 (0.02)	100	2.42	0.002

*Jester-1*⁴. This dataset consists of 24,983 users and 100 jokes, all users have rated a minimum of 36 jokes and ratings are real values from -10 to 10.

Jester-2. This dataset consists of 23,500 users and 100 jokes, all users have rated a minimum of 36 jokes and ratings are real values from -10 to 10.

Jester-3. This dataset consists of 24,938 users and 100 jokes, all users have rated between 15 and 35 jokes and ratings are real values from -10 to 10.

For Movielens we uniformly sampled $\rho = 50\%$ of the available entries for each user for training, as in [Toh and Yun, 2011], for Jester-1 and Jester-2, 20 entries, and for Jester-3, 8 entries, and we used 10% of the these for validation. The error measurement was Normalized Mean Absolute Error as in [Toh and Yun, 2011],

$$\text{NMAE} = \frac{\|\text{true} - \text{predicted}\|^2}{\frac{\#\text{observations}}{r_{\max} - r_{\min}}}.$$

and we used percentage change in objective, with a tolerance of 10^{-3} , averaged over 50 runs. The results are outlined in Table 4. Note that on all three datasets, the spectral k -support performed the best (standard deviations were of the order of 10^{-5}), and we note that our results for trace norm regularization on MovieLens agreed with [Jaggi and Sulovsky, 2010].

⁴<http://goldberg.berkeley.edu/jester-data/>

Table 4: Matrix completion on MovieLens and Jester data sets.

dataset	norm	test error	N	r	k	a
Movie-Lens $\rho = 50\%$	tr	0.2034	57	87	-	-
	en	0.2034	67	87	-	-
	ks	0.2031	67	102	1.00	-
	box	0.2035	61	943	1.00	10^{-5}
Jester1 20 per line	tr	0.1787	14	98	-	-
	en	0.1787	14	98	-	-
	ks	0.1764	14	84	5.00	-
	box	0.1766	16	100	4.00	10^{-6}
Jester2 20 per line	tr	0.1788	14	97	-	-
	en	0.1788	14	97	-	-
	ks	0.1767	13	75	5.00	-
	box	0.1767	14	100	4.00	10^{-6}
Jester3 8 per line	tr	0.1988	16	49	-	-
	en	0.1988	16	49	-	-
	ks	0.1970	17	46	3.70	-
	box	0.1973	23	100	5.91	0.001

6.4 Clustered Multitask Learning

In our final set of experiments we considered the *Lenk personal computer* dataset [Lenk et al., 1996], see also [Argyriou et al., 2008]. This consists of 180 ratings on a scale from 0 to 10 of 20 profiles of computers characterized by 13 features ($d = 14$ with bias term). The clustering is suggested by the assumption that users are motivated by similar groups of features. We used the RMSE of true vs. predicted ratings, normalised over the tasks, with a tolerance of 10^{-5} , averaged over 100 runs.

The results are outlined in Table 5. The cluster norm and centered spectral k -support outperformed the other penalties considerably in all regimes, with the $c-cn$ performing slightly better out of the two. The improvement due to centering compares favourably to similar results for the trace norm in [Argyriou et al., 2008].

7 Conclusion

We considered a family of regularizers and studied a special case of the parameter set. We showed that the k -support norm belongs to this family, and we showed that the matrix extension of the k -support is a special case of the cluster norm for multitask learning. We derived an efficient computation of the norm and an improved computation of the proximity operator of the square norm. Furthermore we presented a method to solve matrix learning problems with the centered cluster norm using proximal gradient methods.

Our empirical results confirm the improved computational performance of our proposed proximity operator algorithm, which enabled us to efficiently solve matrix problems using the

Table 5: Multitask learning clustering on Lenk dataset.

dataset	norm	test error	r	k	a
5 per task	tr	2.2238 (0.17)	9	-	-
	rn	2.1196 (0.06)	11	-	-
	ks	2.1679 (0.07)	10	1.01	-
	box	2.1637 (0.07)	14	1.00	0.0001
	c-ks	1.9309 (0.04)	11	1.01	-
	c-cn	1.9219 (0.05)	14	1.82	0.0006
8 per task	tr	1.9110 (0.05)	14	-	-
	en	1.8983 (0.05)	14	-	-
	ks	1.9158 (0.05)	13	1.04	-
	box	1.9061 (0.05)	14	1.02	0.0003
	c-ks	1.7915 (0.03)	13	2.06	-
	c-cn	1.7912 (0.03)	14	2.06	0.00006
12 per task	tr	1.8702 (0.05)	8	-	-
	en	1.8642 (0.05)	8	-	-
	ks	1.7620 (0.07)	13	1.02	-
	box	1.7194 (0.02)	14	1.01	0.0003
	c-ks	1.6813 (0.02)	14	2.61	-
	c-cn	1.6744 (0.02)	14	1.00	0.0009

spectral penalties outlined herein. Our experiments for matrix completion indicate that the spectral Θ -norm generally outperforms the trace norm and the elastic net. For multitask learning, the centered cluster norm and the centered spectral k -support norm considerably outperformed the trace norm and elastic net. Given the simpler form of the spectral k -support, the results suggest that on datasets exhibiting clustering structure this norm would be a viable alternative to the trace norm.

In future work we would like to derive oracle inequalities and Rademacher complexities for the spectral penalty. It is also of interest to study the broader family of penalties, which admits natural extensions to a general family of operator norms, as well as their relationship to other known families of regularizers.

Acknowledgements

We would like to thank Andreas Argyriou, Andreas Maurer and Charles Micchelli for useful discussions. Part of this work was supported by EPSRC Grant EP/H027203/1 and Royal Society International Joint Project 2012/R2.

A Appendix

In this appendix we provide proofs of some of the results stated in the main body of the paper, and collect some auxiliary results.

Proof of Theorem 2.1. Consider the expression for the dual norm. For fixed θ , the objective f in the supremum is positive, zero only for $u = 0$, absolutely homogeneous of order one and convex. The triangle inequality holds since for such f by homogeneity and convexity we have

$$\begin{aligned} f(u + v) &= 2f\left(\frac{1}{2}u + \frac{1}{2}v\right) \leq 2\frac{1}{2}f(u) + 2\frac{1}{2}f(v) \\ &= f(u) + f(v). \end{aligned}$$

Since the supremum of convex functions is convex, the dual Θ expression is indeed a norm. With respect to the primal expression, we recall the definition of the Fenchel conjugate of $f : \mathbb{R}^d \rightarrow \mathbb{R}$:

$$f^*(u) = \sup_{w \in \mathbb{R}} \langle u, w \rangle - f(w).$$

It is a standard result from convex analysis that for any norm $\|\cdot\|$, the Fenchel conjugate f^* of $f(w) = \frac{1}{2}\|w\|^2$ satisfies $f^*(u) = \frac{1}{2}\|u\|_*^2$, where $\|\cdot\|_*$ is the corresponding dual norm (see, e.g. [Boyd and Vandenberghe, 2004, p. 93]). Furthermore, for any norm, the biconjugate f^{**} is equal to f [Borwein and Lewis, 2000, Th. 4.2.1]. Applying this to the dual norm we have

$$\begin{aligned} \frac{1}{2}\|w\|_\Theta^2 &= f(w) = \sup_u \{\langle w, u \rangle - f^*(u)\} \\ &= \sup_u \left\{ \sum_{i=1}^d w_i u_i - \frac{1}{2} \sup_\theta \sum_{i=1}^d \theta_i u_i^2 \right\} \\ &= \sup_u \inf_\theta \left\{ \sum_{i=1}^d \left(w_i u_i - \frac{1}{2} \theta_i u_i^2 \right) \right\}. \end{aligned}$$

This is a minimax problem in the sense of von Neumann [see e.g. Bertsekas et al., 2003], and we can exchange the order of the inf and the sup, and solve the latter (which is in fact a minimum) componentwise. The gradient with respect to u_i is zero for $u_i = \frac{w_i}{\theta_i}$, and substituting this into the objective we get

$$\frac{1}{2}\|w\|_\Theta^2 = \frac{1}{2} \inf_\theta \sum_{i=1}^d \frac{w_i^2}{\theta_i},$$

and we recover the infimum expression. It follows that eq. (1) defines a norm, and the two norms are duals of each other as required. ■

In order to prove Proposition 2.1, we will prove the result under more general assumptions and derive Proposition 2.1 as a special case. Our proof technique is essentially taken from [Maurer and Pontil, 2012, Theorem 7].

We first show the following.

Lemma A.1. Let $T_n \in S_d$, the set of real valued symmetric $d \times d$ matrices, for $n = 1, \dots, N$, and assume for every $w \in \mathbb{R}^d$, $w \neq 0$, there exists T_n such that $T_n w \neq 0$. Then

$$\max_{n=1}^N \|T_n w\|_2 \quad (21)$$

is a norm and the dual norm is given by

$$\inf_v \left\{ \sum_{n=1}^N \|T_n^+ v_n\|_2 : \sum_{n=1}^N v_n = w, v_n \in \text{Range}(T_n) \right\}$$

where T_n^+ is the pseudo-inverse of T_n .

Proof. We first show that (21) defines a norm. Positivity and homogeneity are immediate. Non-degeneracy ($\|w\| = 0 \Leftrightarrow w = 0$) follows from the assumption that for every $w \in \mathbb{R}^d$, $w \neq 0$, there exists T_n such that $T_n w \neq 0$. To show the triangle inequality, note

$$\begin{aligned} \max_{n=1}^N \|T_n(u+v)\|_2 &\leq \max_{n=1}^N (\|T_n u\|_2 + \|T_n v\|_2) \\ &\leq \max_{n=1}^N \|T_n u\|_2 + \max_{n=1}^N \|T_n v\|_2, \end{aligned}$$

and we are done. To derive the dual norm we need to compute

$$\|w\|_* = \sup_u \{ \langle w, u \rangle : \|T_n^+ u\|_2 \leq 1, n = 1, \dots, N \}.$$

The same assumption implies that $T = \sum_{n=1}^N T_n$ has non trivial nullspace, hence is invertible, and

$$w = T^{-1} \sum_{n=1}^N T_n w = \sum_{n=1}^N T^{-1} T_n w = \sum_{n=1}^N v_n,$$

where $v_n = T^{-1} T_n w$. Letting $v = (v_1, \dots, v_N)$, noting that for $v_n \in \text{Range}(T_n)$, we have $v_n = T_n T_n^+ v_n$, and using symmetry of T_n , we have

$$\begin{aligned} \langle w, u \rangle &= \sum_{n=1}^N \langle T_n T_n^+ v_n, u \rangle \\ &= \sum_{n=1}^N \langle T_n^+ v_n, T_n u \rangle \\ &\leq \sum_{n=1}^N \|T_n^+ v_n\|_2 \|T_n u\|_2 \\ &\leq \sum_{n=1}^N \|T_n^+ v_n\|_2, \end{aligned}$$

where the penultimate inequality follows by Cauchy Schwarz. Furthermore all inequalities are tight for the choice $v_n = T_n^+ u$, if $u \in \arg\max_{n=1}^N \|T_n u\|_2$, and $v_n = 0$ otherwise. The result follows. ■

Note that the assumption that $T_n \in S_d$ is not restrictive since the matrices in eq. (21) inside the Frobenius norm appear in the form $T_n^\top T_n$. For diagonal matrices we have the following corollary of Lemma A.1.

Corollary A.1. *Let $T_n = \text{diag}(t_n)$ for some $t_n \in \mathbb{R}_+^d$ and assume for every $w \in \mathbb{R}^d$, $w \neq 0$, there exist n such that $T_n w \neq 0$. Then*

$$\max_{n=1}^N \|T_n u\|_2$$

is a norm and the dual norm is given by

$$\inf_{v \in \mathcal{V}} \left\{ \sum_{n=1}^N \|T_n^+ v_n\|_2 : \sum_{n=1}^N v_n = w \right\}$$

where $\mathcal{V} = \{(v_n)_{n=1}^N : v_n \in \mathbb{R}^n, \text{supp}(v_n) \subseteq \text{supp}(t_n)\}$ and T_n^+ is the pseudo-inverse of T_n .

We can now prove Proposition 2.1.

Proof of Proposition 2.1. Note that Corollary A.1 applies to the Θ -norm in the case that the closure of Θ has a finite number of extreme points t_1, \dots, t_N . Now observe that the extreme points of the set Θ in equation (2) for the given choice of a, b and c are of the form $1a + 1_g(b - a)$ for $g \in \mathcal{G}_k$, where recall \mathcal{G}_k is the set of all subsets of $\{1, \dots, d\}$ of cardinality k , 1 is the vector of all ones, and (1_g) is the indicator vector for set g . Let $N = |\mathcal{G}_k|$ and identify n with g . Note for this choice that, for $x \in \mathbb{R}^d$

$$\|T_g x\|_2 = \sqrt{\sum_{i \in g} \frac{x_i^2}{b} + \sum_{i \notin g} \frac{x_i^2}{a}},$$

and the proposition is proved. ■

The following result illustrates the relationship between the unit balls relating to $\|T_n \cdot\|$ and the unit ball of $\max_n \|T_n \cdot\|_2$ as defined in eq. (21).

Proposition A.1. *Let $\|\cdot\|$ be the norm defined on \mathbb{R}^d by*

$$\|w\| = \inf_{v \in \mathcal{V}} \left\{ \sum_{n=1}^N \|T_n^+ v_n\| : \sum_{n=1}^N v_n = w, v_n \in \text{Range}(T_n) \right\},$$

for $T_n \in S_d$ with pseudoinverse T_n^+ . Let $B = \{w : \|w\| \leq 1\}$ be the corresponding unit ball, and let $A = \text{co} \left(\bigcup_{n=1}^N A_n \right)$, where

$$A_n = \{w : \|T_n^+ w\| \leq 1, w \in \text{Range}(T_n)\}.$$

Then $B = A$.

Proof. We first show $A \subset B$. Let $w \in A$. By definition of the convex hull, there exist w_n and $\lambda_n \in [0, 1]$ ($n = 1, \dots, N$) such that $w_n \in \text{Range}(T_n)$, $w = \sum_{n=1}^N \lambda_n w_n$, and $\sum_{n=1}^N \lambda_n = 1$. Defining $\hat{v}_n = \lambda_n w_n$, we have $w = \sum_{n=1}^N \hat{v}_n$. It follows that

$$\begin{aligned} \|w\| &= \inf_{\{v_n\}} \left\{ \sum_{n=1}^N \|T_n^+ v_n\| : \sum_{n=1}^N v_n = w, v_n \in \text{Range}(T_n) \right\} \\ &\leq \sum_{n=1}^N \|T_n^+ \hat{v}_n\| = \sum_{n=1}^N \lambda_n \|T_n^+ w_n\| \leq \sum_{n=1}^N \lambda_n = 1. \end{aligned}$$

To show $B \subset A$, let $w \in B$, so $\|w\| \leq 1$. Then there exists a sequence $\{v_1^k, \dots, v_N^k\}$ such that for each k , $\sum_{n=1}^N v_n^k = w$ and $v_n^k \in \text{Range}(T_n)$, and we have

$$\|w\| = \lim_{k \rightarrow \infty} \sum_{n=1}^N \|T_n^+ v_n^k\|.$$

Let $w_n^k = \frac{v_n^k}{\|T_n^+ v_n^k\|}$ and $\lambda_n^k = \|T_n^+ v_n^k\|$. Then for all k , $\sum_{n=1}^N \lambda_n^k = 1$, $\sum_{n=1}^N \lambda_n^k w_n^k = w$, and $w_n^k \in A_n$, and taking the limit we get the desired result. ■

Proof of Theorem 3.1. We solve the constrained optimization problem

$$\inf_{\theta} \left\{ \sum_{i=1}^d \frac{w_i^2}{\theta_i} : a \leq \theta_i \leq b, \sum_{i=1}^d \theta_i \leq c \right\}. \quad (22)$$

To simplify notation we assume without loss of generality that the w_i are ordered non increasing, and it follows by Lemma A.2 that the θ_i are ordered non increasing. We further assume without loss of generality that $w_i \neq 0$ for all i , and $c \leq db$ (see Remark A.1). The objective is continuous and we take the infimum over a closed bounded set, so a solution exists, the solution is a minimum, and it is unique by strict convexity. Furthermore, since $c \leq db$, the sum constraint will be tight at the optimum.

The Lagrangian is given by

$$L(\theta, \alpha) = \sum_{i=1}^d \frac{w_i^2}{\theta_i} + \frac{1}{\alpha^2} \left(\sum_{i=1}^d \theta_i - c \right)$$

where $1/\alpha^2$ is a strictly positive multiplier to be chosen to make the sum constraint tight, call this value α^* . Now let θ^* be the minimizer of $L(\theta, \alpha^*)$ over θ subject to $a \leq \theta_i \leq b$. Then θ^* solves equation (22). To see this, note that we have for any θ , $L(\theta^*, \alpha^*) \leq L(\theta, \alpha^*)$, so we have

$$\sum_{i=1}^d \frac{w_i^2}{\theta_i^*} \leq \sum_{i=1}^d \frac{w_i^2}{\theta_i} + \frac{1}{(\alpha^*)^2} \left(\sum_{i=1}^d \theta_i - c \right).$$

As this holds for all $\theta \in [a, b]^d$, then it holds for all such θ where additionally the constraint $\sum_{i=1}^d \theta_i \leq c$ holds. In this case the second term on the right hand side is at most zero, so we have for all such θ

$$\sum_{i=1}^d \frac{w_i^2}{\theta_i^*} \leq \sum_{i=1}^d \frac{w_i^2}{\theta_i},$$

whence it follows that θ^* is the minimizer of (11).

We can therefore solve the original problem by minimizing the Lagrangian over the box constraint. Due to the coupling effect of the multiplier, the problem is separable, and we can solve the simplified problem componentwise using Lemma A.3. It follows that

$$\theta_i = \begin{cases} a, & \text{if } \alpha < \frac{a}{|w_i|}, \\ \alpha |w_i|, & \text{if } \frac{a}{|w_i|} \leq \alpha \leq \frac{b}{|w_i|}, \\ b, & \text{if } \alpha > \frac{b}{|w_i|}, \end{cases} \quad (23)$$

where $\alpha > 0$ is such that $\sum_{i=1}^d \theta_i(\alpha) = c$. The minimizer then has the form

$$\theta = (\underbrace{b, \dots, b}_q, \theta_{q+1}, \dots, \theta_{d-\ell}, \underbrace{a, \dots, a}_\ell)$$

where $q, \ell \in \{0, \dots, d\}$ are determined by the value of α , which satisfies

$$S(\alpha) = \sum_{i=1}^d \theta_i(\alpha) = qb + \sum_{i=q+1}^{d-\ell} \alpha |w_i| + \ell a = c,$$

i.e. $\alpha = \frac{p}{\sum_{i=q+1}^{d-\ell} |w_i|}$ where $p = c - qb - \ell a$. The value of the norm follows by substituting θ into the objective and we get

$$\begin{aligned} \|w\|_\Theta^2 &= \sum_{i=1}^q \frac{|w_i|^2}{b^2} + \frac{1}{p} \left(\sum_{i=q+1}^{d-\ell} |w_i| \right)^2 + \sum_{i=d-\ell+1}^d \frac{|w_i|^2}{a^2} \\ &= \frac{1}{b^2} \|w_Q\|_2^2 + \frac{1}{p} \|w_I\|_1^2 + \frac{1}{a^2} \|w_L\|_2^2, \end{aligned}$$

as required. We can further characterise q and ℓ by considering the form of θ_i . By construction we have $\theta_q \geq b > \theta_{q+1}$ and $\theta_{d-\ell} > a \geq \theta_{d-\ell+1}$, or equivalently

$$\begin{aligned} \frac{|w_q|}{b} &\geq \frac{1}{p} \sum_{i=q+1}^{d-\ell} |w_i| > \frac{|w_{q+1}|}{b} \\ \frac{|w_{d-\ell}|}{a} &\geq \frac{1}{p} \sum_{i=q+1}^{d-\ell} |w_i| > \frac{|w_{d-\ell+1}|}{a}, \end{aligned}$$

as required. ■

Remark A.1. The case where some w_i are zero follows from the case that we have considered in the theorem. If $w_i = 0$ for $n < i \leq d$, then clearly we must have $\theta_i = a$ for all such i . We then consider the n -dimensional problem of finding $(\theta_1, \dots, \theta_n)$ that minimizes $\sum_{i=1}^n \frac{w_i^2}{\theta_i}$, subject to $a \leq \theta_i \leq b$, and $\sum_{i=1}^n \theta_i \leq c'$, where $c' = c - (d - n)a$. As $c \geq da$ by assumption, we also have $c' < na$, so a solution exists to the n dimensional problem. If $c' < bn$, then a solution is trivially $\theta_i = b$ for all $i = 1 \dots n$. In general, $c' \geq bn$, and we proceed as per the proof of the theorem. Finally, a θ that solves the original d -dimensional problem will be given by $(\theta_1, \dots, \theta_n, a, \dots, a)$.

The following auxiliary result is used in Theorem 3.1.

Lemma A.2. Let $w_1 \geq \dots \geq w_d > 0$, $\Theta = \{\theta : a \leq \theta_i \leq b, \sum_{i=1}^d \theta_i \leq c\}$ ($0 < a < b$), and let $\theta^* = \operatorname{argmin}_{\theta \in \Theta} \sum_{i=1}^d \frac{w_i^2}{\theta_i}$. Then $\theta_1^* \geq \dots \geq \theta_d^*$.

Proof. Let $i < j$, so $w_i \geq w_j$, and suppose that $\theta_i^* < \theta_j^*$. Define $\hat{\theta}$ to have identical elements to θ^* , except with the i and j elements exchanged. As θ^* is the unique argmin of a strictly convex function, comparing the objectives we have

$$\begin{aligned} \frac{w_i^2}{\theta_i^*} + \frac{w_j^2}{\theta_j^*} &< \frac{w_i^2}{\hat{\theta}_i} + \frac{w_j^2}{\hat{\theta}_j} = \frac{w_i^2}{\theta_j^*} + \frac{w_j^2}{\theta_i^*} \\ w_i^2 \left(\frac{1}{\theta_i^*} - \frac{1}{\theta_j^*} \right) &< w_j^2 \left(\frac{1}{\theta_i^*} - \frac{1}{\theta_j^*} \right). \end{aligned}$$

Dividing by $\frac{1}{\theta_i^*} - \frac{1}{\theta_j^*} > 0$ we get $w_i < w_j$, a contradiction. ■

We further make use of the following result, which follows from [Micchelli et al., 2013, Theorem 3.1].

Lemma A.3. Let $0 < a < b$, and let $f(\theta) = \sum_{i=1}^d \left(\frac{w_i^2}{\theta_i} + \beta^2 \theta_i \right)$ and consider $\min_{a \leq \theta_i \leq b} f(\theta)$. Then the minimizer $\theta \in \mathbb{R}^d$ is given by

$$\theta_i = \begin{cases} a, & \text{if } \frac{|w_i|}{\beta} < a, \\ \frac{|w_i|}{\beta}, & \text{if } a \leq \frac{|w_i|}{\beta} \leq b, \\ b, & \text{if } \frac{|w_i|}{\beta} > b. \end{cases}$$

Proof. For fixed w , the objective function is strictly convex on \mathbb{R}_{++}^d and has a unique minimum on $(0, \infty)$ (see Figure 1.b in [Micchelli et al., 2013] for a one-dimensional illustration). The problem is separable and we solve

$$\min_{a \leq \theta_i \leq b} \frac{w_i^2}{\theta_i} + \beta^2,$$

which has gradient with respect to θ_i given by $-\frac{w_i^2}{\theta_i^2} + \beta$. This is zero for $\theta_i = \frac{|w_i|}{\beta} =: \theta_i^*$, strictly negative below θ_i^* and strictly increasing above θ_i^* . Considering these three cases we recover the expression in statement of the lemma. ■

References

- A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.
- A. Argyriou, C. A. Micchelli, M. Pontil, L. Shen, and Y. Xu. Efficient first order methods for linear composite regularizers. *CoRR*, abs/1104.1436, 2011.
- A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems 25*, pages 1466–1474, 2012.
- A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009.
- D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, 2003.
- J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization*. Springer-Verlag, 2000.
- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- J.-F. Cai, E. J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2008.
- P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inv Prob.* Springer, 2011.
- T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *J. of Machine Learning Research*, 6:615–637, 2005.
- R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.
- L. Jacob, F. Bach, and J.-P. Vert. Clustered multi-task learning: a convex formulation. *Advances in Neural Information Processing Systems (NIPS 21)*, 2009a.
- L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. *Proc of the 26th Int. Conf. on Machine Learning*, 2009b.
- M Jaggi and M. Sulovsky. A simple algorithm for nuclear norm regularized problems. *Proceedings of the 27th International Conference on Machine Learning*, 2010.
- P. J. Lenk, W. S DeSarbo, P. E. Green, and M. R. Young. Hierarchical bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2):173–191, 1996.
- A. Maurer and M. Pontil. Structured sparsity and generalization. *The Journal of Machine Learning Research*, 13:671–690, 2012.

- R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.
- C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.
- C. A. Micchelli, J. M. Morales, and M. Pontil. Regularizers for structured sparsity. *Advances in Comp. Mathematics*, 38:455–489, 2013.
- Y. Nesterov. Gradient methods for minimizing composite objective function. *Center for Operations Research and Econometrics*, 76, 2007.
- G. Obozinski and F. Bach. Convex relaxation for combinatorial penalties. *CoRR*, 2012.
- K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *SIAM Journal on Imaging Sciences*, 4:573–596, 2011.
- J. Von Neumann. *Some matrix-inequalities and metrization of matrix-space*. Tomsk. Univ. Rev. Vol I, 1937.
- M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(I):49–67, 2006.
- H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.